Published online 6 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D767–D772***
doi:10.1093/nar/gkn892

Human Protein Reference Database—2009 update

T. S. Keshava Prasad[1], Renu Goel[1], Kumaran Kandasamy[1,2,3,4,5], Shivakumar Keerthikumar[1,2], Sameer Kumar[1,2], Suresh Mathivanan[1,2], Deepthi Telikicherla[1,2], Rajesh Raju[1,2], Beema Shafreen[1], Abhilash Venugopal[1,2], Lavanya Balakrishnan[1], Arivusudar Marimuthu[1,3,4,5], Sutopa Banerjee[1], Devi S. Somanathan[1], Aimy Sebastian[1], Sandhya Rani[1], Somak Ray[1], C. J. Harrys Kishore[1], Sashi Kanth[1], Mukhtar Ahmed[1], Manoj K. Kashyap[1,2,3,4,5], Riaz Mohmood[2], Y. L. Ramachandra[2], V. Krishna[2], B. Abdul Rahiman[2], Sujatha Mohan[1], Prathibha Ranganathan[1], Subhashri Ramabadran[1], Raghothama Chaerkady[1,3,4,5] and Akhilesh Pandey[3,4,5,*]

[1]Institute of Bioinformatics, International Tech Park, Bangalore 560 066, [2]Department of Biotechnology, Kuvempu University, Shankaraghatta, Karnataka, India, [3]McKusick-Nathans Institute of Genetic Medicine, [4]Department of Biological Chemistry and [5]Department of Pathology and Oncology, Johns Hopkins University, Baltimore, MD 21205, USA

Received September 16, 2008; Revised October 20, 2008; Accepted October 22, 2008

ABSTRACT

Human Protein Reference Database (HPRD—http://www.hprd.org/), initially described in 2003, is a database of curated proteomic information pertaining to human proteins. We have recently added a number of new features in HPRD. These include PhosphoMotif Finder, which allows users to find the presence of over 320 experimentally verified phosphorylation motifs in proteins of interest. Another new feature is a protein distributed annotation system—Human Proteinpedia (http://www.humanproteinpedia.org/)—through which laboratories can submit their data, which is mapped onto protein entries in HPRD. Over 75 laboratories involved in proteomics research have already participated in this effort by submitting data for over 15 000 human proteins. The submitted data includes mass spectrometry and protein microarray-derived data, among other data types. Finally, HPRD is also linked to a compendium of human signaling pathways developed by our group, NetPath (http://www.netpath.org/), which currently contains annotations for several cancer and immune signaling pathways. Since the last update, more than 5500 new protein sequences have been added, making HPRD a comprehensive resource for studying the human proteome.

INTRODUCTION

Human Protein Reference Database (HPRD; http://www.hprd.org/) is a resource for experimentally derived information about the human proteome including protein–protein interactions, post-translational modifications (PTMs) and tissue expression (1–4). The contents of several proteomic databases, including HPRD, pertaining to human proteins have recently been evaluated in terms of the number of nonredundant protein–protein interactions, number of direct interactions per protein, number of proteins with disease annotation and the number of linked citations (5). The curation and annotation process in HPRD involves entry of protein data through BioBuilder, a tool developed by our group for editing and managing data through a web browser (6). We have incorporated new features, such as PhosphoMotif Finder, links to a signaling pathway resource called NetPath, Human Proteinpedia for enhanced community participation and the use of BLAST for querying mRNA/protein data. Since the last update, we have added approximately 5500 new protein sequences and corresponding information in HPRD, which now contains information on most of the human proteins including their isoforms.

*To whom correspondence should be addressed. Tel: +410 502 6662; Fax: +410 502 7544; Email: pandey@jhmi.edu
Correspondence may also be addressed to T. S. Keshava Prasad. Tel: (+91) 80-28416140; Fax: (+91) 80-28416132; Email: keshav@ibioinformatics.org

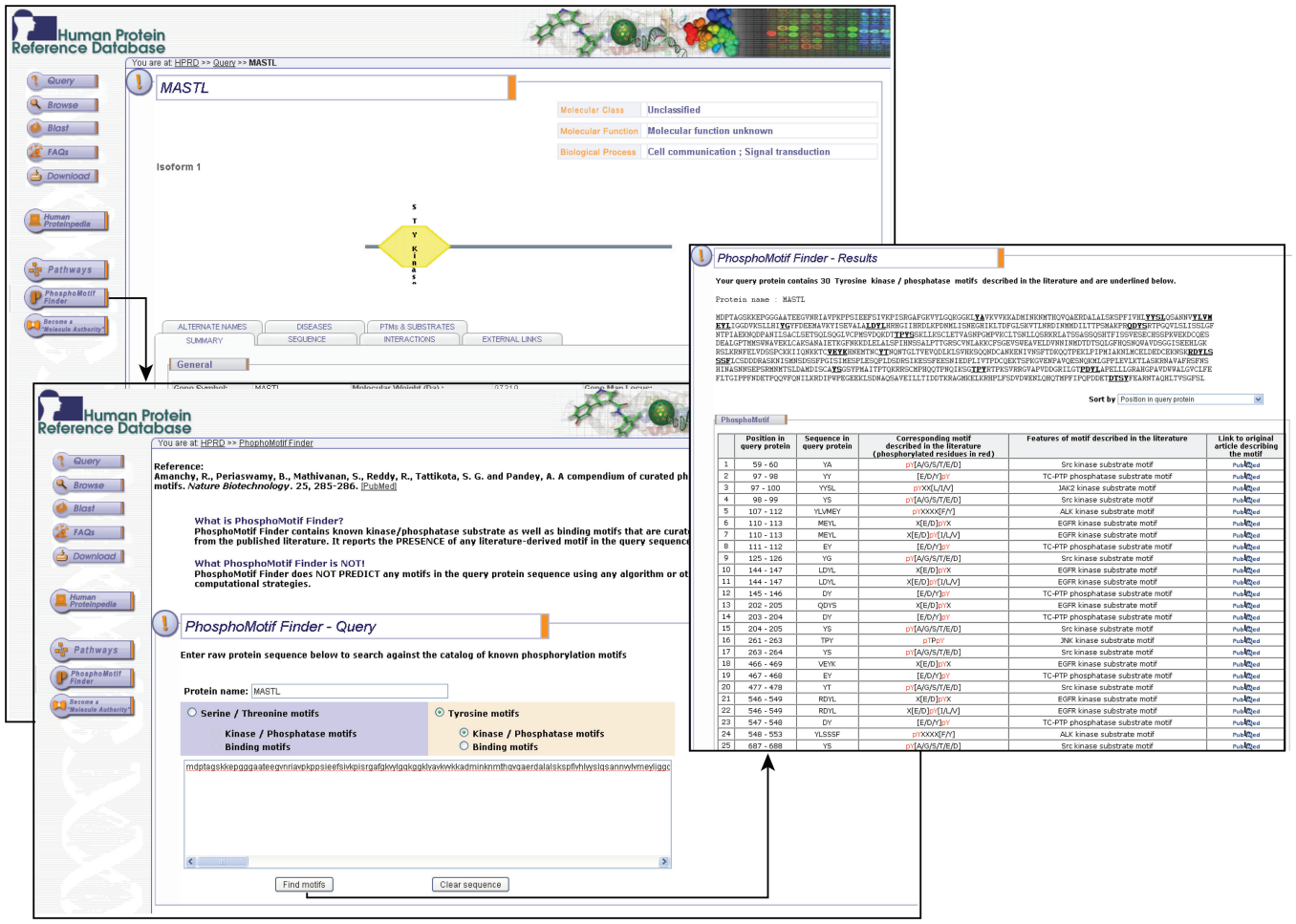

Figure 1. Display of PhosphoMotif Finder integrated into HPRD. Screen shot shows molecule page of MASTL, a hypothetical protein implicated in autosomal dominant thrombocytopenia. 'PhosphoMotif Finder' tab in the HPRD page leads to the utility page where the sequence of the MASTL is displayed. Users can select either serine/threonine or tyrosine motifs and submit the query by clicking 'Find Motifs' button. Result page displays mapped experimentally derived motifs present in sequence along with the information on position, actual sequence, experimentally derived consensus phosphorylation motifs and link to the PubMed abstracts where these motifs have been described. MASTL sequence is shown to contain 30 potential tyrosine phosphorylation sites as seen in this figure.

'PhosphoMotif Finder' searches experimentally derived phosphorylation-based substrate and binding motifs

PhosphoMotif Finder contains experimentally characterized phosphorylation-based substrate and binding motifs derived from the literature (7) and has been integrated with HPRD. PhosphoMotif Finder searches across the user submitted protein sequence for the presence of any of the 320 phosphorylation-based motifs listed in the compendium. Figure 1 shows the presence of 30 known tyrosine kinase phosphorylation sites in microtubule-associated serine/threonine kinase-like protein (MASTL), which is implicated in thrombocytopenia, a blood disorder. In addition to the mapped motifs, PhosphoMotif Finder also indicates potential enzymes (i.e. kinases or phosphatases) associated with these phosphorylation motifs. PhosphoMotif Finder should also be helpful in ascertaining the novelty of any motif that is described in the literature. Finally, it can be used in designing phosphorylation motif-specific antibodies and antibody-based arrays.

'NetPath' pathway resource

We have incorporated a compendium of human signaling pathways called NetPath (http://www.netpath.org/) through the 'Pathways' tab in HPRD. NetPath contains information about protein interactions, catalytic reactions and protein translocation events, which occur downstream of ligand–receptor interactions. Currently, the role of 2732 and 1793 proteins are thus annotated in the context of cancer and immune signaling pathways, respectively. We have also cataloged genes that are upregulated or downregulated at the transcriptional level under the influence of these signaling pathways. Pathway data can be downloaded in standard international data exchange formats including BioPAX Level 2.0, PSI-MI version 2.5 and SBML version 2.1. The list of transcriptionally upregulated and downregulated genes can be obtained in the

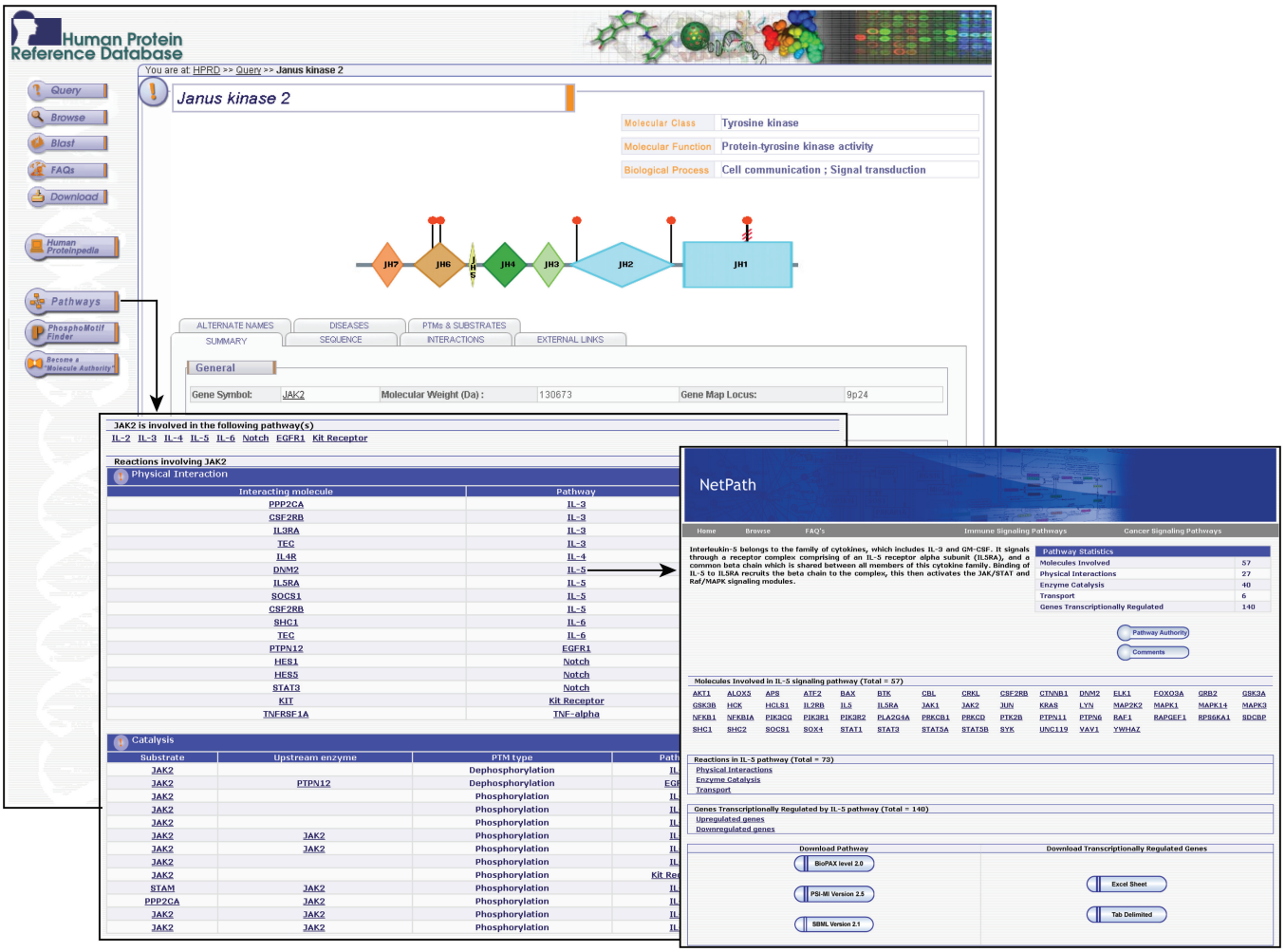

Figure 2. Linking to human signaling pathways from HPRD. 'Pathways' button in the HPRD page of JAK2 is hyperlinked to its NetPath page. It shows the list of signaling pathways in which the protein is involved along with the description of its interactors in each pathway. Each interaction or catalysis event is linked to the PubMed abstract of the original article. The pathway name is linked to the specific signaling pathway annotated in NetPath.

form of Excel sheet and tab delimited text documents. Integration of NetPath data in HPRD will assist users in visualizing the probable role of proteins in diverse signaling networks. For example, Janus Kinase 2 (JAK2) is involved in diverse pathways including EGFR1, Kit receptor, Notch, IL-2, IL-3, IL-4, IL-5 and IL-6 signaling pathways. NetPath provides the list of physical interactions and catalysis events of JAK2 with various proteins under different signaling pathways. Each interaction or catalysis event is linked to the PubMed abstract of the original article (Figure 2).

Annotation of proteomic information

Protein isoforms. We have included most of human protein isoforms present in the RefSeq Database (8). Currently, 25 661 protein sequences encoded by 19 433 genes have been annotated in HPRD. Phosphodiesterase 9A, cAMP response element modulator, collagen type XIII alpha1 and dystrophin are examples of proteins with the highest number of isoforms with 20, 20, 19 and 18 isoforms, respectively. However, only data pertaining to the sequence, subcellular localization, mRNA/protein expression, biological motifs and domains are currently being annotated as isoform specific whereas protein–protein interactions and enzyme–substrate relationships are annotated as common to all isoforms. This is mainly due to the general lack of experimental data for the latter.

Protein–protein interactions. Protein–protein interactions are one of the most requested components of HPRD among those who downloaded this dataset. We have added more than 5000 protein–protein interactions in HPRD since the previous update in 2006. Among the 38 167 protein–protein interactions documented in HPRD, 8958 interactions were based on yeast two-hybrid analysis alone, whereas 8827 interactions were based on *in vitro* and 7163 on *in vivo* methods. Detection of 2410 protein–protein interactions was confirmed by all three methods. Overall, in HPRD, 8710 proteins are annotated with at least one protein–protein interaction, whereas 2015 and 774 proteins have more than

Table 1. Statistics of proteomic data annotated by HPRD team and submitted to Human Proteinpedia

Dataset	Dataset annotated by HPRD team	Data submitted through Human Proteinpedia
Protein–protein interactions	38 167	15 231
PTMs	16 972	17 410
Subcellular localization	19 670	2906
mRNA/protein expression	65 536	150 368

Table 2. Statistics of PTM data annotated among various PTM types

PTM type	Count
Phosphorylation	10 858
Dephosphorylation	3118
Glycosylation	1860
Sumolylation	305
Acetylation	259
Methylation	274
Palmitoylation	149
Myristoylation	43
Glutathionylation	11
ADP-ribosylation	7
Others	88
Total	16 972

5 or 10 protein–protein interactions, respectively. The 14-3-3 gamma protein has a maximum of 173 protein–protein interactions. 15 231 protein–protein interactions (Table 1) have been submitted to HPRD by the scientific community using Human Proteinpedia (9,10). Enzyme–substrate relationships determined through peptide/protein arrays is a new data type included in HPRD, as represented by the phosphorylation of Tyr 16 of RNA binding motif protein 10 by c-Src.

PTMs and subcellular localization. HPRD currently contains information for 16 972 PTMs (Table 2) which belong to various categories with phosphorylation (10 858), dephosphorylation (3118) and glycosylation (1860) forming the majority of the annotated PTMs (Table 2). At least one enzyme responsible for PTMs has been annotated for 8960 PTMs, which resulted in the documentation of 7253 enzyme–substrate relationships. Of these, 1277 PTMs have more than one enzyme annotated. Human Proteinpedia has contributed over 17 400 PTMs, which are mainly derived from mass spectrometry studies. One or more site of subcellular localization has been annotated for 8620 proteins in HPRD with 586 of them being isoform specific. In addition to these, scientific investigators have contributed 2906 entries pertaining to subcellular localization through Human Proteinpedia.

Community participation through 'Human Proteinpedia'

We have developed a distributed annotation system called Human Proteinpedia and incorporated in HPRD (9,10). Proteomic investigators can directly contribute protein data derived from diverse platforms including the yeast two-hybrid, mass spectrometry, peptide/protein array, immunohistochemistry, Western blot, coimmunoprecipitation and fluorescence microscopy to HPRD using Human Proteinpedia. The protein features that can be mapped to corresponding entries in HPRD include PTMs, mRNA/protein expression in tissues or cell lines, subcellular localization, enzyme–substrate relationships and protein–protein interactions. These annotations are made available for viewing in a separate box beneath the HPRD annotation (Figure 3). Each entry is also linked to experimental evidence, such as mass spectra, images of Western blots and fluorescence micrographs. Figure 3 shows five serine phosphorylation sites for Adducin 1 protein in HPRD, submitted through Human Proteinpedia. PTM sites are linked to the meta-annotation of mass spectrometry data in Human Proteinpedia database as submitted by the investigator. The corresponding MS/MS spectrum can also be viewed by following a link in the meta-annotation page.

Investigators worldwide have already submitted 15 231 protein–protein interactions, 17 410 PTMs and 150 368 mRNA/protein expression to HPRD through Human Proteinpedia. Human Proteinpedia has increased quantity of the HPRD data by 2-fold in a relatively short span of time (Table 1). By involving investigators and experimentalists in the annotation of proteomic data, Human Proteinpedia has transformed HPRD into a true community database.

Usage of HPRD data by the community

Over the years, the biomedical community has provided valuable suggestions by interacting with HPRD team through 'Comments' and 'Help' buttons provided in HPRD page. More than 8000 gene comments, expert suggestions and help requests have been received and nearly 100 scientists have been designated as 'Molecule Authorities' based on their expertise. We hope to further increase participation by the community by implementing a microattribution system, which provides a citable credit to the investigators. Web resources that display or have made use of HPRD data include Entrez-Gene, VisANT (11) Genes2Networks (12), Cerebral (13), BioNetBuilder (14), COXPRESdb (15), STRING 7 (16) and UniHI (17). Molecular Signature Database (MSigDB) (18) used for Gene Set Enrichment Analysis of gene expression data incorporates pathway gene sets curated from HPRD. Sequence analysis tools which use HPRD data include CompariMotif (19) and SLiMFinder (20). CutDB, a database of proteolytic events (21), PepBank, a database of peptides (22) and T1Dbase, a database for type 1 diabetes research (23) are other resources that also incorporate curated proteomic data from HPRD.

CONCLUSIONS

With the inclusion of most of human protein sequences, HPRD has grown into an integrated knowledgebase for genomic and proteomic investigators. Incorporation of PhosphoMotif Finder and signaling pathways will help

Figure 3. Display of PTM data in HPRD submitted through Human Proteinpedia. Adducin1 molecule page in HPRD shows five novel phosphorylation sites submitted through Human Proteinpedia. Phosphorylation sites are hyperlinked to Human Proteinpedia page with information on the investigator, laboratory and meta-annotation of mass spectrometry experiment. Corresponding MS/MS spectrum for a peptide is also displayed using spectrum viewer developed by PRIDE.

users to generate novel hypotheses or to point out likely molecules involved in a biological process of their interest. Further, the implementation of Human Proteinpedia has transformed HPRD into a community driven database and we hope that this trend will continue so that each and every entry is directly or indirectly verified by the individual experimentalists.

ACKNOWLEDGEMENTS

We thank all investigators and 'Molecule Authorities' who have provided valuable feedback about individual entries in this database.

FUNDING

Funding for open access charge: Institute of Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Gandhi,T.K., Zhong,J., Mathivanan,S., Karthick,L., Chandrika,K.N., Mohan,S.S., Sharma,S., Pinkert,S., Nagaraju,S., Periaswamy,B. *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.*, **38**, 285–293.
2. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
3. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
4. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, **32**, D497–D501.
5. Mathivanan,S., Periaswamy,B., Gandhi,T.K., Kandasamy,K., Suresh,S., Mohmood,R., Ramachandra,Y.L. and Pandey,A. (2006) An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, **7(Suppl. 5)**, S19.

6. Navarro,J.D., Talreja,N., Peri,S., Vrushabendra,B.M., Rashmi,B.P., Padma,N., Surendranath,V., Jonnalagadda,C.K., Kousthub,P.S., Deshpande,N., Shanker,K. *et al*. (2004) BioBuilder as a database development and functional annotation platform for proteins. *BMC Bioinformatics*, **20**, 5–43.
7. Amanchy,R., Periaswamy,B., Mathivanan,S., Reddy,R., Tattikota,S.G. and Pandey,A. (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.*, **25**, 285–286.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al*. (2008) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **36**, D13–D21.
9. Kandasamy,K., Keerthikumar,S., Goel,R., Mathivanan,S., Patankar,N., Shafreen,B., Renuse,S., Pawar,H., Ramachandra,Y.L., Acharya,P.K. *et al*. (2008) Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res.* (in press).
10. Mathivanan,S., Ahmed,M., Ahn,N.G., Alexandre,H., Amanchy,R., Andrews,P.C., Bader,J.S., Balgley,B.M., Bantscheff,M., Bennett,K.L. *et al*. (2008) Human Proteinpedia enables sharing of human protein data. *Nat. Biotechnol.*, **26**, 164–167.
11. Hu,Z., Snitkin,E.S. and DeLisi,C. (2008) VisANT: an integrative framework for networks in systems biology. *Brief Bioinform.*, **9**, 317–325.
12. Berger,S.I., Posner,J.M. and Ma'ayan,A. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, **8**, 372.
13. Barsky,A., Gardy,J.L., Hancock,R.E. and Munzner,T. (2007) Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, **23**, 1040–1042.
14. Avila-Campillo,I., Drew,K., Lin,J., Reiss,D.J. and Bonneau,R. (2007) BioNetBuilder: automatic integration of biological networks. *Bioinformatics*, **23**, 392–393.
15. Obayashi,T., Hayashi,S., Shibaoka,M., Saeki,M., Ohta,H. and Kinoshita,K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, **36**, D77–D82.
16. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
17. Chaurasia,G., Iqbal,Y., Hanig,C., Herzel,H., Wanker,E.E. and Futschik,M.E. (2007) UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res.*, **35**, D590–D594.
18. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
19. Edwards,R.J., Davey,N.E. and Shields,D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics*, **24**, 1307–1309.
20. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.
21. Igarashi,Y., Eroshkin,A., Gramatikova,S., Gramatikoff,K., Zhang,Y., Smith,J.W., Osterman,A.L. and Godzik,A. (2007) CutDB: a proteolytic event database. *Nucleic Acids Res.*, **35**, D546–D549.
22. Shtatland,T., Guettler,D., Kossodo,M., Pivovarov,M. and Weissleder,R. (2007) PepBank–a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics*, **8**, 280.
23. Hulbert,E.M., Smink,L.J., Adlem,E.C., Allen,J.E., Burdick,D.B., Burren,O.S., Cassen,V.M., Cavnor,C.C., Dolman,G.E., Flamez,D. *et al*. (2007) T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res.*, **35**, D742–D746.